Estimation Non-Paramétrique de Densités

Machine Learning

Cédric RICHARD
Université Côte d'Azur

Estimation non-paramétrique de densité

Objectifs

Nous avons supposé que les fonctions de densité $p(\boldsymbol{x}|\omega_i)$ sont connues :

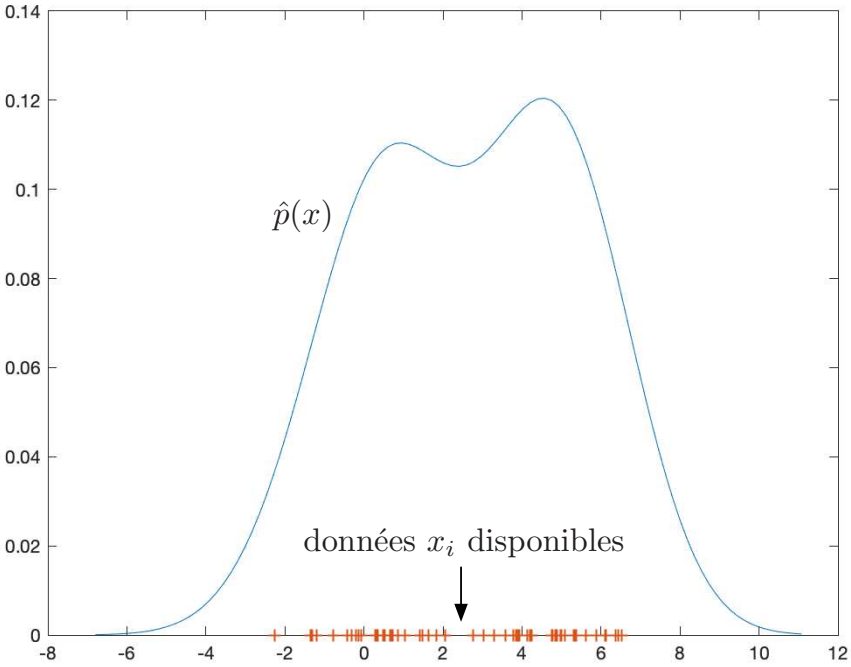
- ▷ soit parfaitement, par leur expression et les paramètres associés
  exemple : loi normale $\mathcal{N}(\mu, \sigma^2)$ de moyenne μ et de variance σ^2 connues
- ▷ soit partiellement par leur expression, de paramètres inconnus à estimer
  exemple : loi normale $\mathcal{N}(\mu, \sigma^2)$ de moyenne μ et de variance σ^2 à estimer

On ne connait généralement pas les fonctions de densité de probabilité. Il est alors nécessaire de les inférer à partir des données disponibles

Estimation non-paramétrique de densité

Exemple (1D)

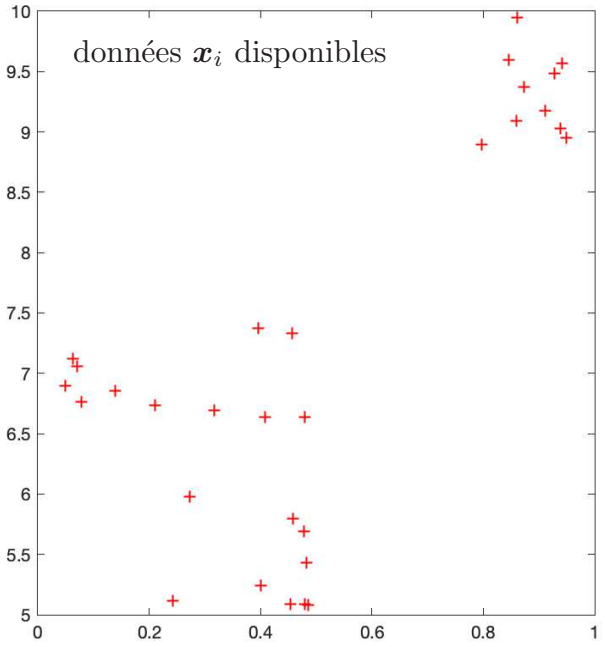
Estimation non-paramétrique de densité attendue :

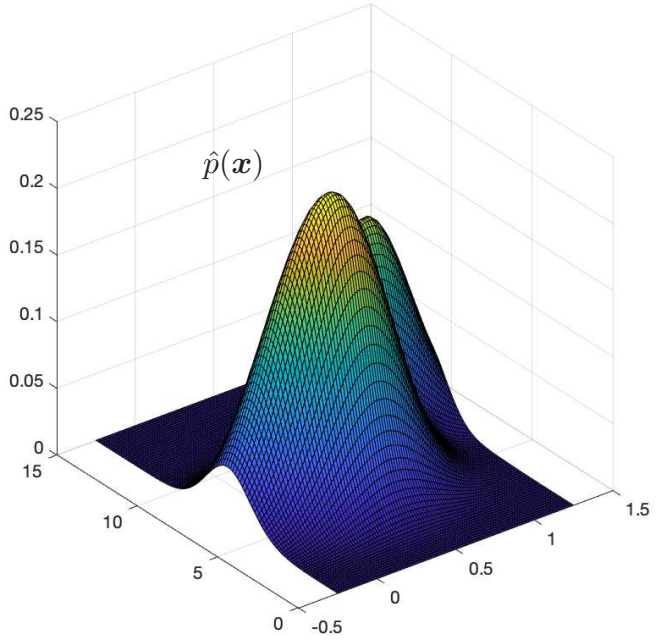


Estimation non-paramétrique de densité

Exemple (2D)

Estimation non-paramétrique de densité attendue :





Estimation non-paramétrique de densité

Histogramme

Une méthode élémentaire est l'histogramme :

- ▷ on subdivise l'espace des observations en boîtes
- ▷ on approxime la densité au centre de chaque boîte par la fraction du nombre de données d'apprentissage en chaque boîte

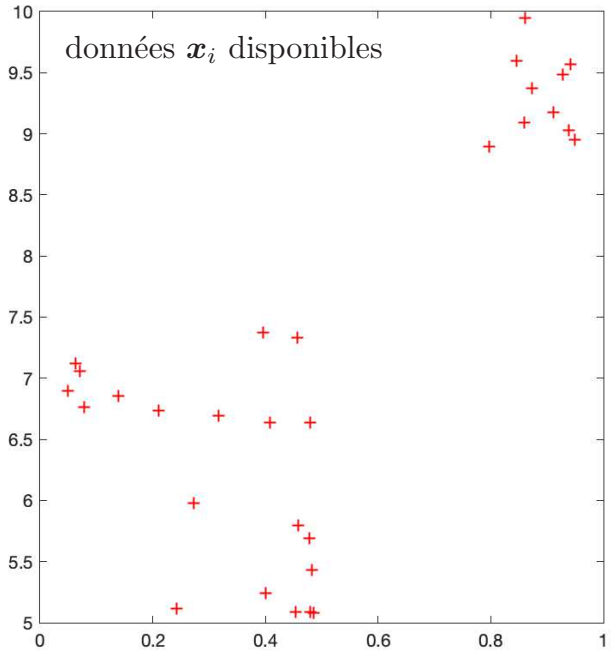
$$\hat{p}(\boldsymbol{x}) = \frac{1}{N} \frac{\#\text{ de données } \boldsymbol{x}_k \text{ dans la même boîte que } \boldsymbol{x}}{\text{volume de la boîte}}$$

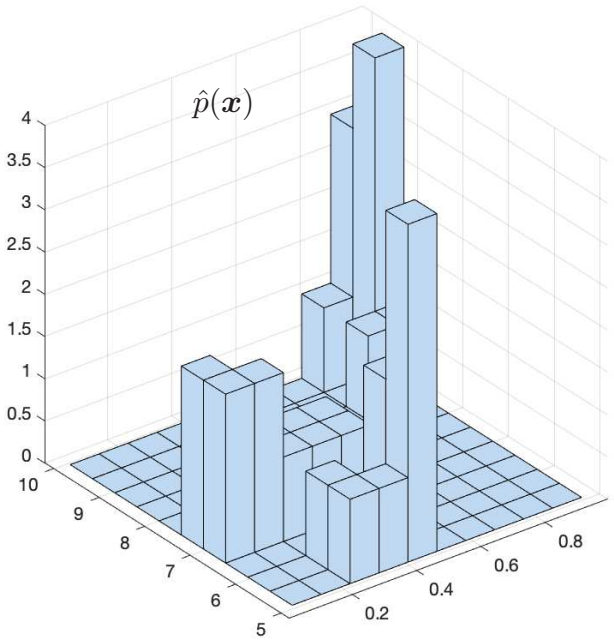
Deux paramètres nécessitent d'être fixés :

- ▷ position de la première boîte
- ▷ dimensions des boîtes

Estimation non-paramétrique de densité

Exemple d'histogramme



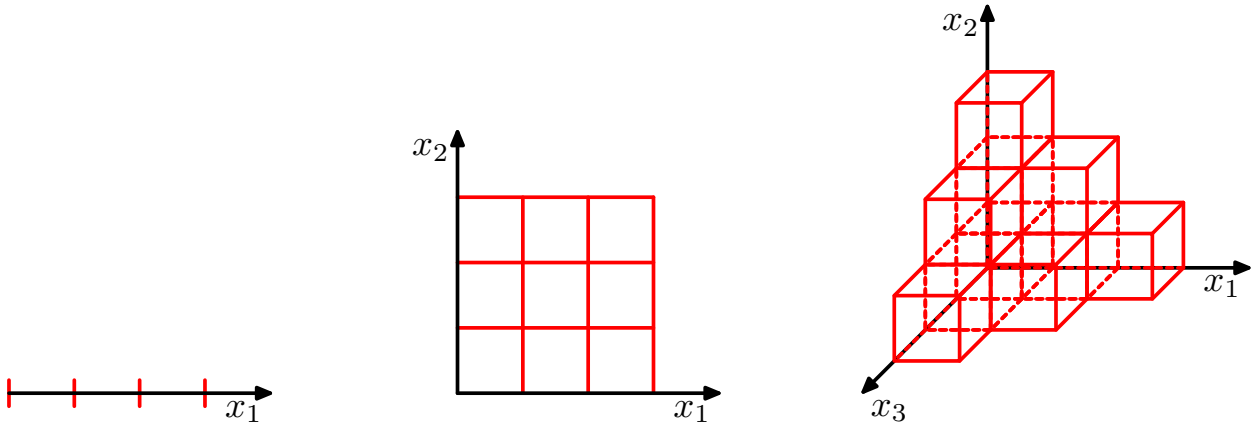


Estimation non-paramétrique de densité

Histogramme

Les histogrammes ne sont pas des solutions satisfaisantes, hormis pour une visualisation 1D ou 2D simple :

- ▷ la densité estimée dépend de la position des boîtes et de leur orientation
- ▷ la densité estimée n'est pas lisse
- ▷ la malédiction de la dimensionnalité, due au nombre exponentiel de boîtes nécessaires en fonction de la dimension des données



Estimation non-paramétrique de densité

Principes généraux

▷ La probabilité que $\boldsymbol{x}$, distribué selon une loi $p(\boldsymbol{x})$, appartient à une région $\mathcal{R}$ de l'espace des observations est :

$$P = \int_{\boldsymbol{x} \in \mathcal{R}} p(\boldsymbol{x}) d\boldsymbol{x}$$

▷ Soit N observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ distribuées selon $p(\boldsymbol{x})$. La probabilité que k d'entre elles appartiennent à $\mathcal{R}$ est donnée par la distribution binomiale :

$$P(k) = C_N^k P^k (1-P)^{N-k}$$

Rappels :

$$E[k] = NP \quad \text{var}[k] = NP(1-P)$$

▷ En conséquence, on a :

$$E\left[\frac{k}{N}\right] = P \quad \text{var}\left[\frac{k}{N}\right] = \frac{P(1-P)}{N}$$

Le ratio des données dans $\mathcal{R}$ est donc une bonne estimée de P : $P \approx \frac{k}{N}$

Estimation non-paramétrique de densité

Principes généraux

- Le ratio des données dans $\mathcal{R}$ est une bonne estimée de P : $P \approx \frac{k}{N}$
- En supposant que $\mathcal{R}$ est suffisamment petite pour que $p(\boldsymbol{x})$ n'y varie pas de façon significative, on peut écrire

$$P = \int_{\boldsymbol{x}' \in \mathcal{R}} p(\boldsymbol{x}') d\boldsymbol{x}' \approx p(\boldsymbol{x}) V$$

  où V est le volume de $\mathcal{R}$
- En combinant ces 2 résultats, on aboutit à

$$\boxed{p(\boldsymbol{x}) \approx \frac{k}{NV}}$$

Cette estimée est d'autant meilleure que N est grand et que V est petit.
En pratique, N est fixé et il faut trouver un compromis entre N et V.

Estimation non-paramétrique de densité

Principes généraux

Les méthodes d'estimation non-paramétriques de densité reposent sur :

$$\boxed{p(\boldsymbol{x}) \approx \frac{k}{NV}}$$

Deux stratégies coexistent :

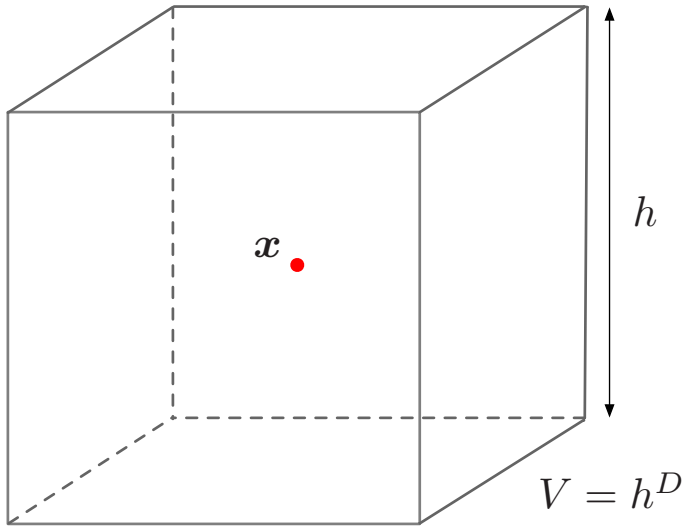
- ▷ on fixe V et on détermine k : méthodes d'estimation à noyau
- ▷ on fixe k et on détermine V : méthode des k-plus-proches-voisins (kPPV)

Ces 2 stratégies font successivement l'objet de ce cours

Estimation non-paramétrique de densité

Fenêtre de Parzen

On suppose que la région $\mathcal{R}$ enfermant les données est un hypercube dans $\mathbb{R}^D$, de côté de longueur h, centré en $\boldsymbol{x}$.



Estimation non-paramétrique de densité

Fenêtre de Parzen

▷ Afin de déterminer le nombre de données contenues dans le cube $\mathcal{R}$, on introduit la fonction noyau

$$K(\boldsymbol{u}) = \begin{cases} 1 & |u_j| < 1/2 \qquad \forall j = 1, \ldots, D \\ 0 & \text{sinon} \end{cases}$$

Il s'agit d'un hypercube de longueur de côté 1 centré sur l'origine

▷ On a donc

$$K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right) = \begin{cases} 1 & \text{si } \boldsymbol{x}_n \in \mathcal{R} \\ 0 & \text{sinon} \end{cases}$$

▷ Le nombre total k de données situées dans $\mathcal{R}$ est

$$\boxed{k = \sum_{n=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)}$$

Estimation non-paramétrique de densité

Estimateur de Parzen

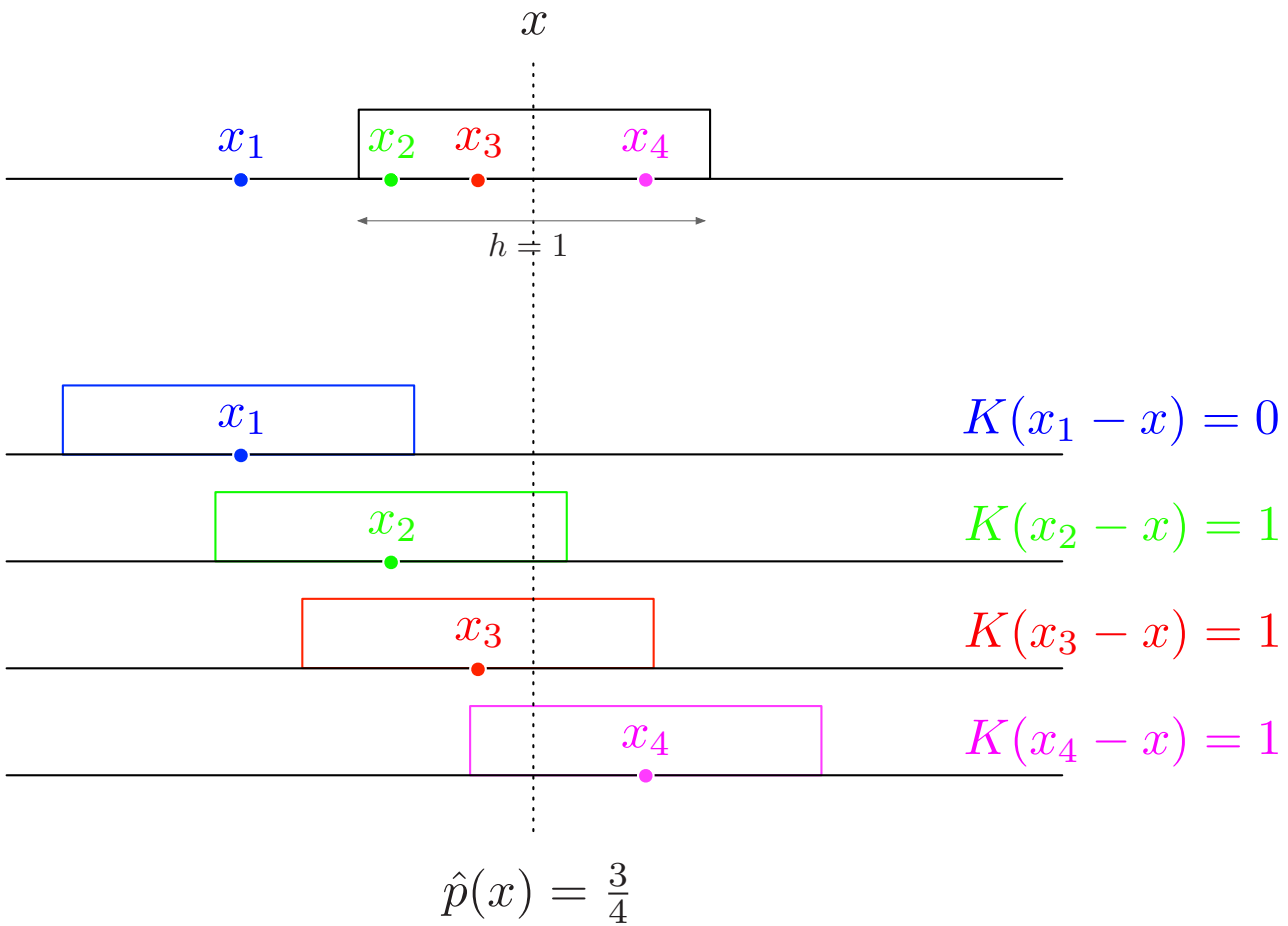
L'estimateur de Parzen au point $\boldsymbol{x}$ est défini par

$$\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^D} \sum_{n=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)$$

Remarque :
Contrairement aux histogrammes, les boîtes de l'estimateur de Parzen sont centrées sur les données

Estimation non-paramétrique de densité

Estimateur de Parzen



Estimation non-paramétrique de densité

Estimateur de Parzen

On calcule l'espérance mathématique de $\hat{p}(\boldsymbol{x})$:

$$\begin{aligned} E[\hat{p}(\boldsymbol{x})] &= \frac{1}{Nh^D} \sum_{n=1}^{N} E\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)\right] \\ &= \frac{1}{h^D} E\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)\right] \\ &= \frac{1}{h^D} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{h}\right) p(\boldsymbol{x}')d\boldsymbol{x}' \end{aligned}$$

où les données $\boldsymbol{x}_n$ sont supposées indépendantes et distribuées selon $p(\boldsymbol{x})$.
En conséquence :

$$\boxed{E[\hat{p}(\boldsymbol{x})] = p(\boldsymbol{x}) * K(\boldsymbol{x}/h)}$$

où $*$ est le produit de convolution
Pour $K(\boldsymbol{x}/h) = \delta(\boldsymbol{x})$, on a $E[\hat{p}(\boldsymbol{x})] = p(\boldsymbol{x})$

Estimation non-paramétrique de densité

Estimateurs à noyau

La fenêtre rectangulaire a plusieurs inconvénients :

- ▷ la densité estimée a des discontinuités
- ▷ toutes les données $\boldsymbol{x}_i$ dans le cube cube ont le même poids, quelle que soit leur distance au centre $\boldsymbol{x}$

Afin de pallier ces inconvénients, d'autres fenêtres $K(\boldsymbol{u})$ existent. Elles sont non-négatives $K(\boldsymbol{u}) \geq 0$, symétriques $K(\boldsymbol{u}) = K(-\boldsymbol{u})$ et vérifient :

$$\int_{\mathbb{R}^D} K(\boldsymbol{u})\, d\boldsymbol{u} = 1$$

Il en résulte l'estimateur :

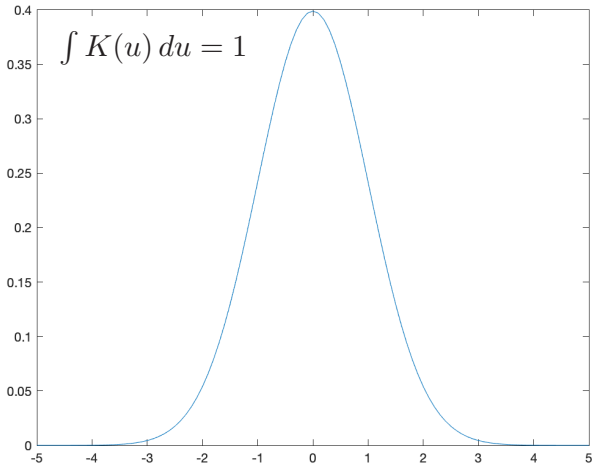
$$\boxed{\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^D} \sum_{n=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)}$$

Estimation non-paramétrique de densité

Exemple d'estimateur à noyau

La fenêtre gaussienne :

$$K(\boldsymbol{u}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\frac{\|\boldsymbol{u}\|^2}{2}\right)$$

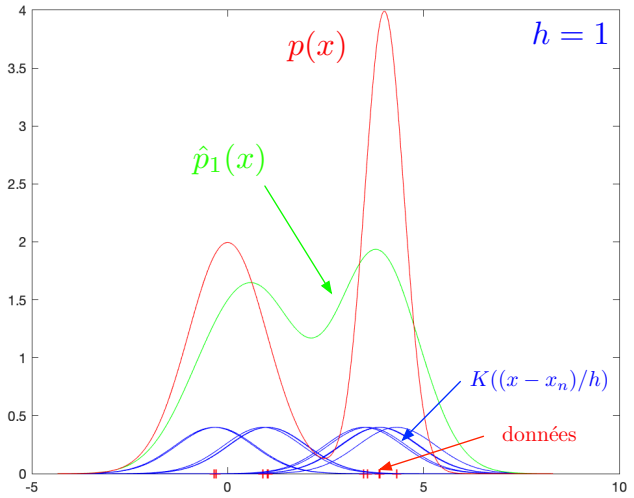


Estimation non-paramétrique de densité

Estimateurs à noyau

$$\hat{p}(\boldsymbol{x}) = \frac{1}{Nh^D} \sum_{n=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_n}{h}\right)$$

- ▷ la densité estimée est une somme de motifs centrés sur les données
- ▷ la fonction noyau $K(\boldsymbol{u})$ détermine la forme des motifs
- ▷ la largeur de bande h, ou paramètre de lissage, définie la largeur des motifs

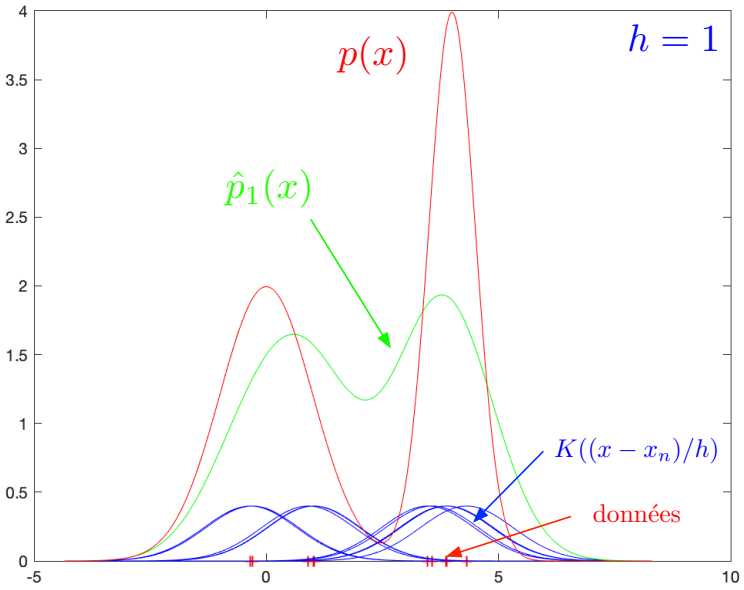


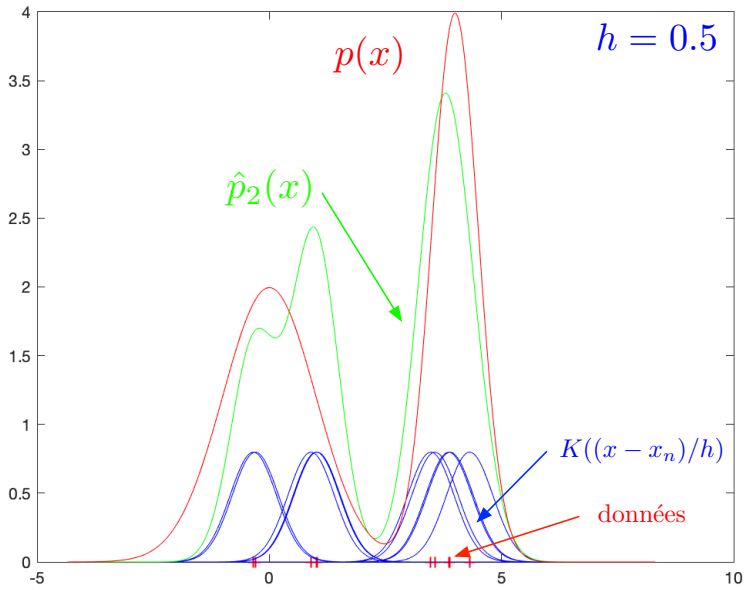
Estimation non-paramétrique de densité

Largeur de bande h

Le choix de la largeur de bande h est crucial :

- ▷ $h \nearrow$: densité estimée $\hat{p}(\boldsymbol{x})$ lisse, masquant éventuellement les modes de $p(\boldsymbol{x})$
- ▷ $h \searrow$: densité estimée $\hat{p}(\boldsymbol{x})$ piquée, potentiellement difficile à interpréter





Estimation non-paramétrique de densité

Estimation de la largeur de bande h

Le choix subjectif de h n'est pas forcément possible, lorsque la dimension D des données est grande.
Il existe des méthodes automatiques pour le faire, reposant elles-mêmes sur des hypothèses pas nécessairement vérifiées

Exemple :

- ▷ On fait l'hypothèse d'une distribution standard $p(\boldsymbol{x})$, et on recherche h optimum au sens de l'erreur quadratique moyenne :

$$\hat{h} = \arg\min_h E\left[\int \left[p(\boldsymbol{x}) - \hat{p}(\boldsymbol{x})\right]^2 d\boldsymbol{x}\right]$$

- ▷ En choisissant une loi normale pour $p(\boldsymbol{x})$, on trouve

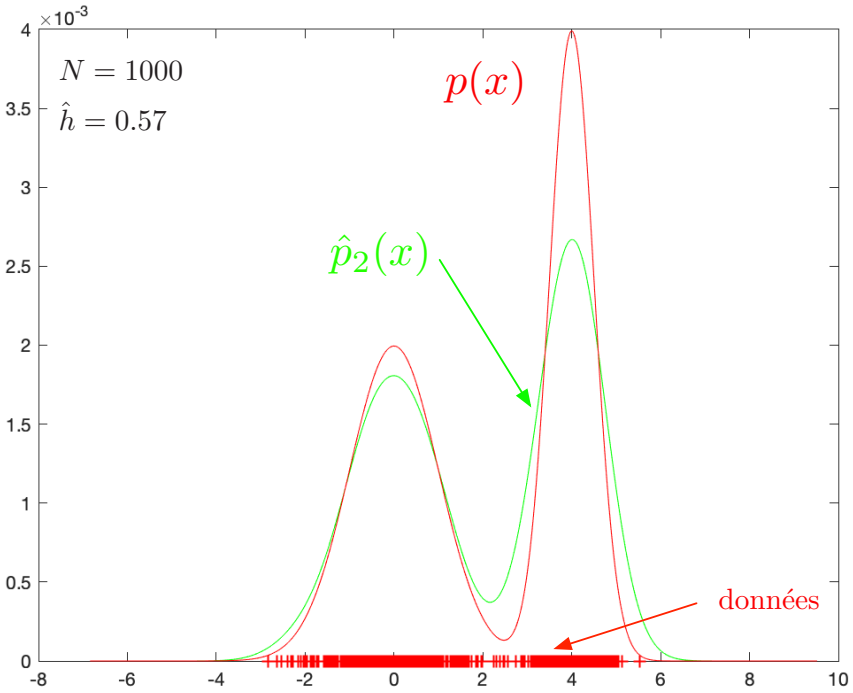
$$\hat{h} = 1.06\,\hat{\sigma} N^{-1/5}$$

où $\hat{\sigma}$ est l'écart-type estimé à partir des données disponibles

Estimation non-paramétrique de densité

Estimation de la largeur de bande h

Par la méthode précédente ($h^* = 1.06\,\hat{\sigma} N^{-1/5}$) :



Estimation non-paramétrique de densité

Estimation de densités multi-variées

Les densités multi-variées, i.e., $p(\boldsymbol{x})$ avec $\boldsymbol{x} \in \mathbb{R}^D$, peuvent être estimées à l'aide de noyaux scalaires, i.e., $K(u)$ avec $u \in \mathbb{R}$

Méthode : produit de noyaux = noyau produit

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} K(\boldsymbol{x}; \boldsymbol{x}_n, h_1, \ldots, h_D)$$

$$\text{avec } K(\boldsymbol{x}; \boldsymbol{x}_n, h_1, \ldots, h_D) = \frac{1}{h_1 \ldots h_D} \prod_{d=1}^{D} K_d \left(\frac{x_d - x_{n,d}}{h_d} \right)$$

où

$K_d(u)$ est le noyau scalaire associé à la dimension d

h_d est la largeur de bande associée au noyau $K_d(u)$

$x_{n,d}$ est la d^{e} composante de $\boldsymbol{x}_n$

Estimation non-paramétrique de densité

Estimation de densités multi-variées

L'usage d'un noyau produit n'induit pas que les variables de $\boldsymbol{x}$ sont indépendantes

Si les variables étaient indépendantes, on pourrait écrire :

$$p(\boldsymbol{x}) = \prod_{d=1}^{D} p_d(x_d)$$

que l'on estimerait par :

$$\begin{aligned}\hat{p}(\boldsymbol{x}) &= \prod_{d=1}^{D} \hat{p}_d(x_d) \\ &= \prod_{d=1}^{D} \left(\frac{1}{N h_d} \sum_{n=1}^{N} K\left(\frac{x_d - x_{n,d}}{h_d} \right) \right)\end{aligned}$$

Estimation non-paramétrique de densité

Conclusions sur les estimateurs à noyau

Avantages :

- ▷ applicable à des données issues de n'importe quelle distribution $p(\boldsymbol{x})$
- ▷ convergence asymptotique de $\hat{p}(\boldsymbol{x})$ vers $p(\boldsymbol{x})$ pour $N \to +\infty$

Inconvénients :

- ▷ nécessite de disposer d'un grand nombre de données
- ▷ choix de la largeur de bande h de la fenêtre non trivial
- ▷ nécessite de disposer de moyens de calcul performants

Estimation non-paramétrique de densité

k-plus-proches-voisins

Les méthodes d'estimation non-paramétriques de densité reposent sur :

$$\boxed{p(\boldsymbol{x}) \approx \frac{k}{NV}}$$

où

$$\begin{cases} k & \text{nombre de données voisines de } \boldsymbol{x} \\ V & \text{volume du voisinage de } \boldsymbol{x} \\ N & \text{nombre total de données} \end{cases}$$

Deux stratégies coexistent :

- ▷ on fixe V et on détermine k : méthodes d'estimation à noyau
- ▷ on fixe k et on détermine V : méthode des k-plus-proches-voisins (kPPV)

Estimation non-paramétrique de densité

k-plus-proches-voisins

Principe :

- ▷ On fait croître le volume du voisinage de $\boldsymbol{x}$ jusqu'à ce qu'il inclut k voisins
- ▷ On évalue le volume V du voisinage de $\boldsymbol{x}$

$$V = c_D \, r_k^D(\boldsymbol{x}) \text{ avec } c_D = \frac{\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2}+1\right)}$$

où

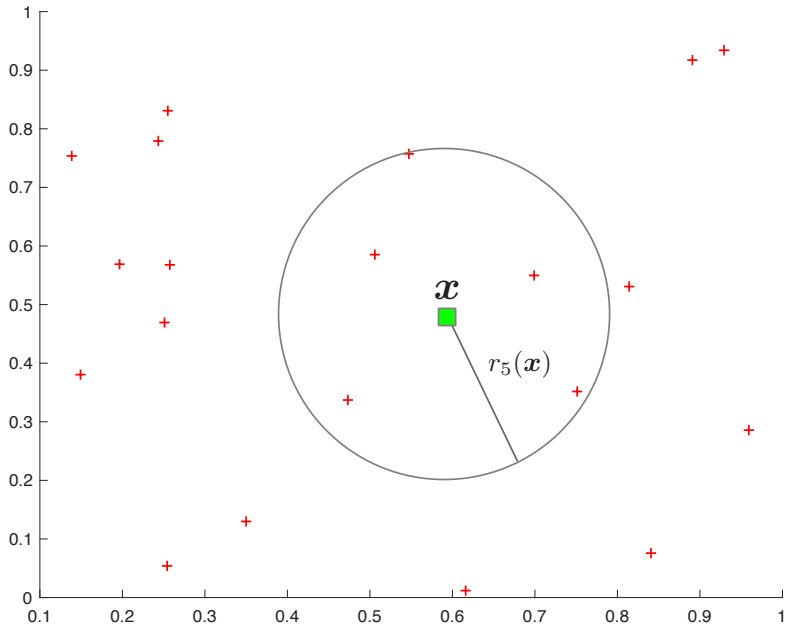
$r_k(\boldsymbol{x})$ est la distance euclidienne entre $\boldsymbol{x}$ et son k^{e} plus proche voisin

$\Gamma(n+1) = n\Gamma(n)$ avec $\Gamma(1/2) = \sqrt{\pi}$ et $\Gamma(1) = 1$

D	$\Gamma\left(\frac{D}{2}+1\right)$	c_D	V
1	$\frac{1}{2}\sqrt{\pi}$	2	$2r_k$
2	1	π	πr_k^2
3	$\frac{3}{4}\sqrt{\pi}$	$\frac{4}{3}\pi$	$\frac{4}{3}\pi r_k^3$

Estimation non-paramétrique de densité

k-plus-proches-voisins



Estimation non-paramétrique de densité

k-plus-proches-voisins

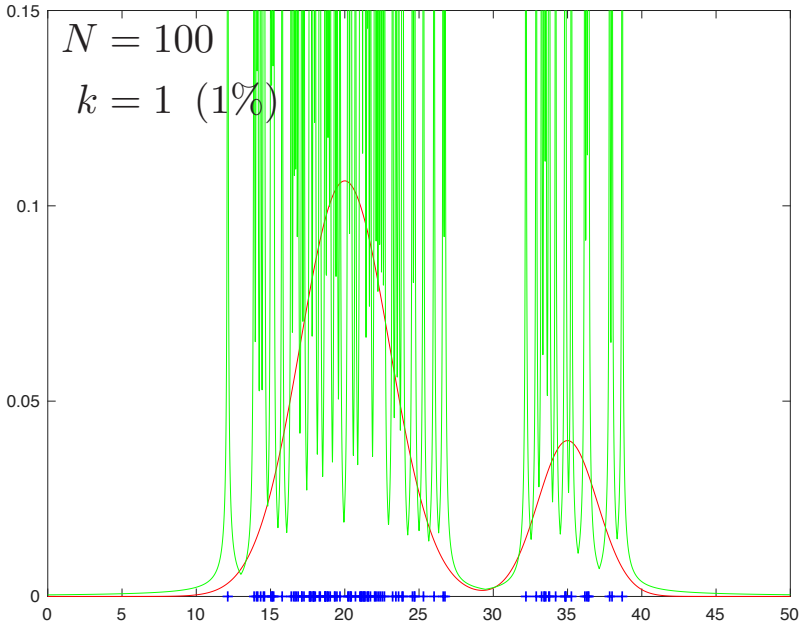
L'estimation de densité par k-plus-proches-voisins n'est pas satisfaisante car

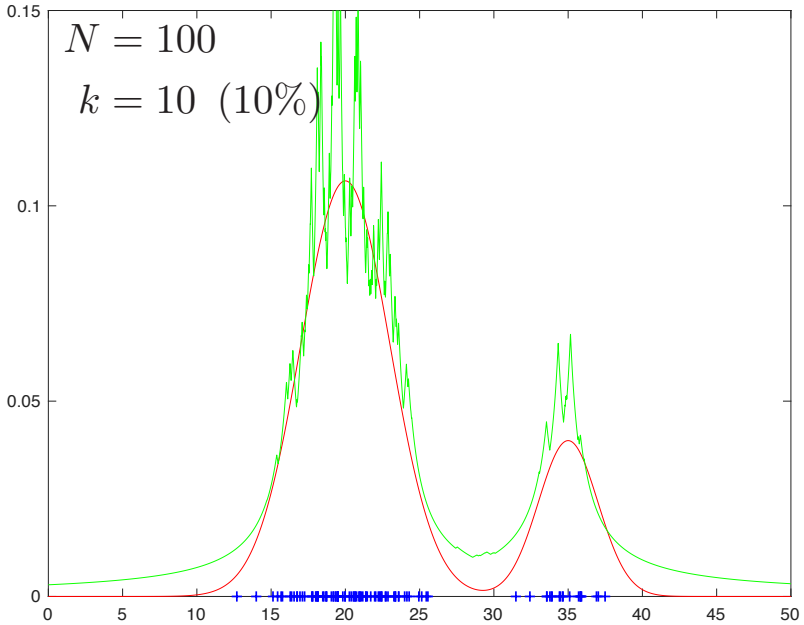
- ▷ les estimées sont sujettes au bruit
- ▷ les estimées présentes des queues lourdes
- ▷ les estimées présentes des discontinuités
- ▷ les estimées ne sont pas des densités de probabilité car leur intégrale sur l'ensemble des observations diverge

Estimation non-paramétrique de densité

k-plus-proches-voisins : exemple 1D

Illustration avec $p(x) = 0.8\,\mathcal{N}(20, 9) + 0.2\,\mathcal{N}(35, 4)$

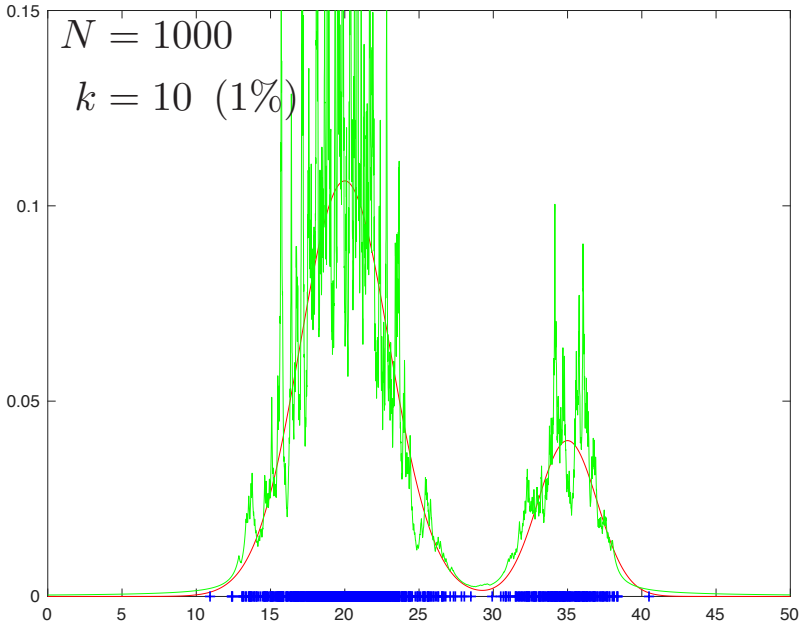


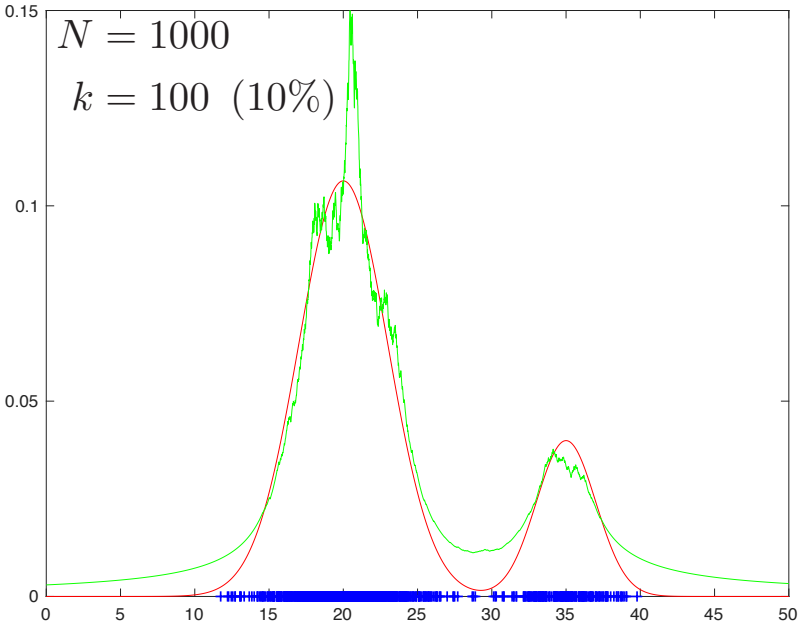


Estimation non-paramétrique de densité

k-plus-proches-voisins : exemple 1D

Illustration avec $p(x) = 0.8\,\mathcal{N}(20, 9) + 0.2\,\mathcal{N}(35, 4)$

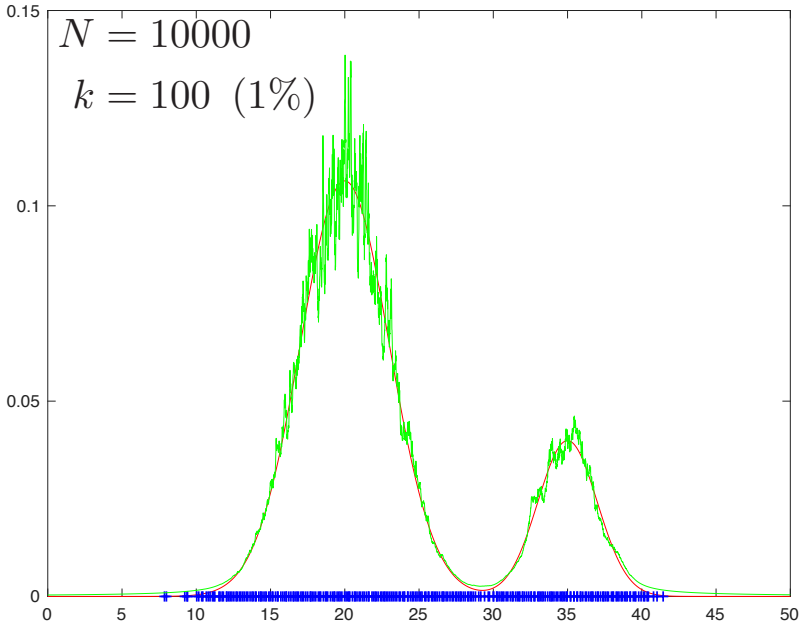


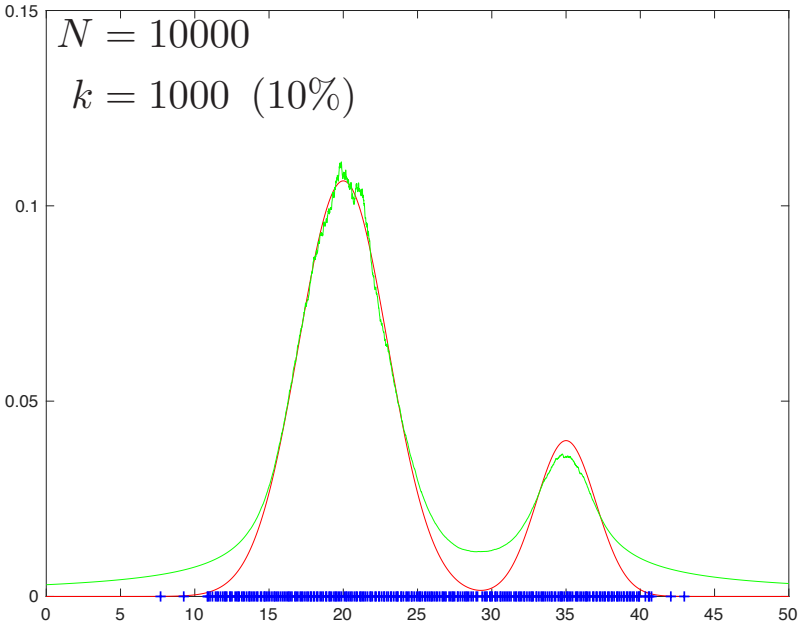


Estimation non-paramétrique de densité

k-plus-proches-voisins : exemple 1D

Illustration avec $p(x) = 0.8\,\mathcal{N}(20, 9) + 0.2\,\mathcal{N}(35, 4)$

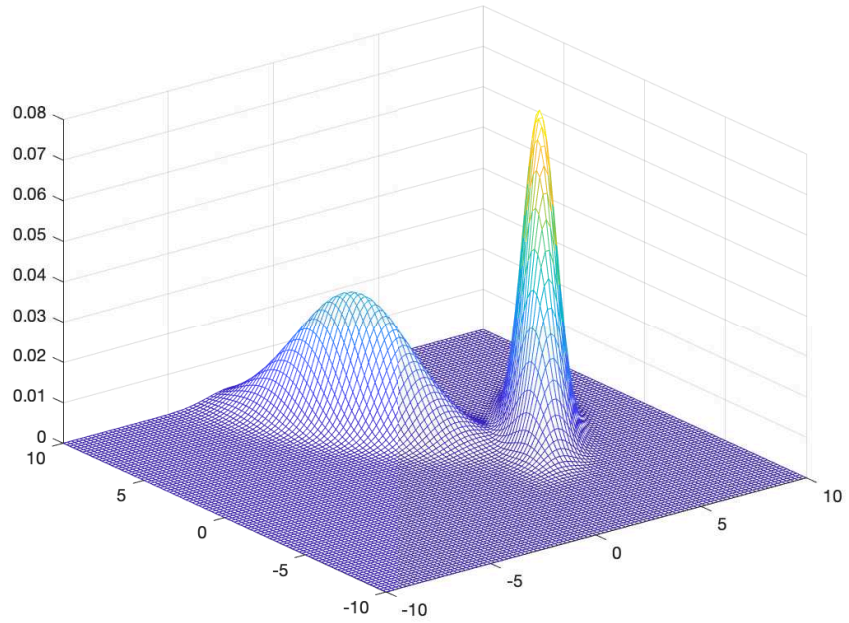




Estimation non-paramétrique de densité

k-plus-proches-voisins : exemple 2D

Illustration avec $p(x) = 0.5\,\mathcal{N}\left(\begin{bmatrix}0\\5\end{bmatrix}, \begin{bmatrix}2 & -2\\-2 & 5\end{bmatrix}\right) + 0.5\,\mathcal{N}\left(\begin{bmatrix}5\\0\end{bmatrix}, \begin{bmatrix}1 & 1\\1 & 2\end{bmatrix}\right)$



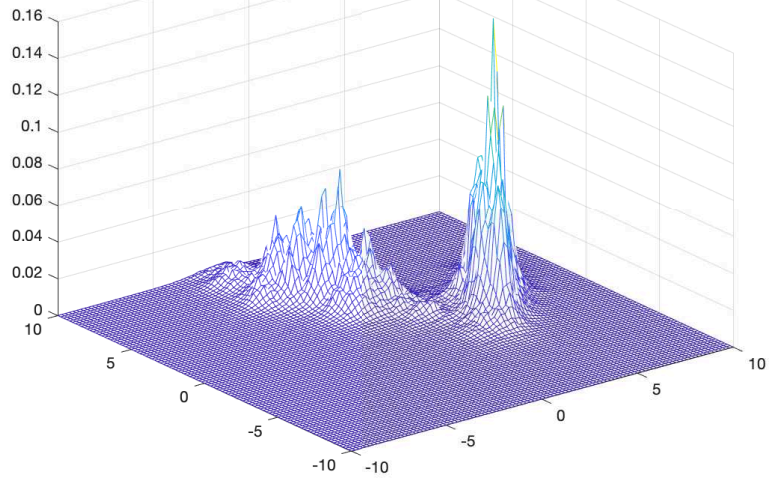
Estimation non-paramétrique de densité

k-plus-proches-voisins : exemple 2D

Illustration avec $p(x) = 0.5\,\mathcal{N}\left(\begin{bmatrix}0\\5\end{bmatrix}, \begin{bmatrix}2 & -2\\-2 & 5\end{bmatrix}\right) + 0.5\,\mathcal{N}\left(\begin{bmatrix}5\\0\end{bmatrix}, \begin{bmatrix}1 & 1\\1 & 2\end{bmatrix}\right)$

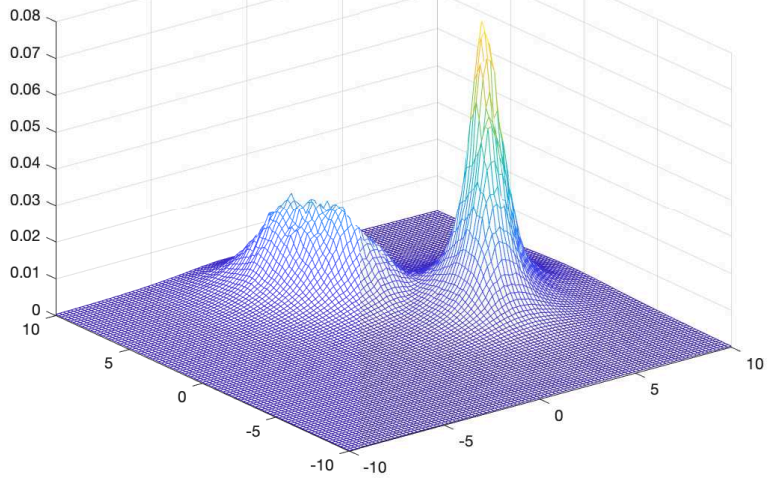
$N = 1000$

$k = 10$ (1%)



$N = 1000$

$k = 100$ (10%)



Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

L'estimation de densité par k-plus-proches-voisins n'est pas satisfaisante. Toutefois, cette méthode offre une approximation simple du classifieur optimum de Bayes.

Problème : N données d'apprentissage, divisées en N_i données par classe ω_i

On estime

- ▷ les fonctions de vraisemblance par : $p(\boldsymbol{x}|\omega_i) = \frac{k_i}{N_i V}$ avec V le voisinage de $\boldsymbol{x}$
- ▷ la fonction de densité de probabilité : $p(\boldsymbol{x}) = \frac{k}{NV}$
- ▷ les probabilités *a priori* : $P(\omega_i) = \frac{N_i}{N}$

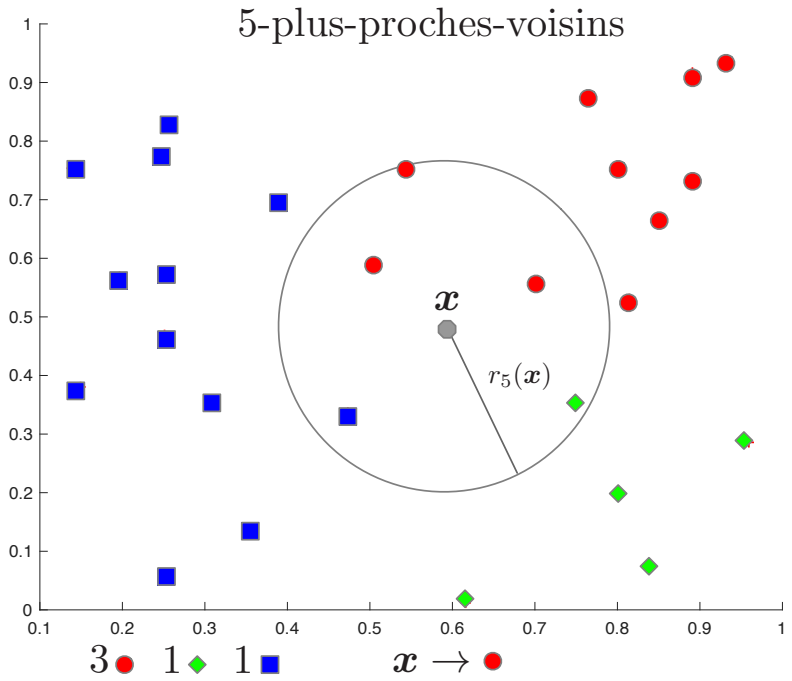
Les fonctions de densité de probabilité *a posteriori* sont données par :

$$\boxed{p(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} = \frac{\frac{k_i}{N_i V}\frac{N_i}{N}}{\frac{k}{NV}} = \frac{k_i}{k}}$$

Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

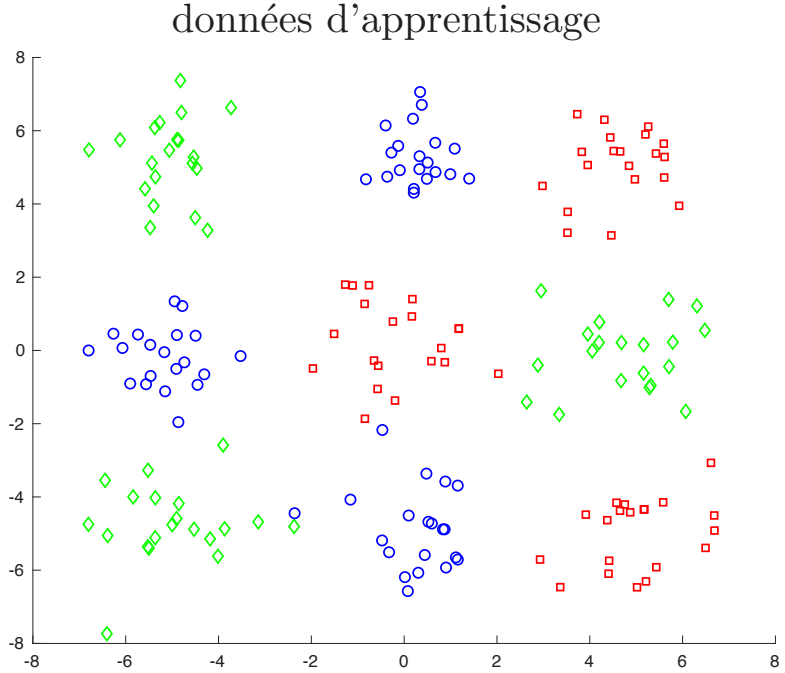
La méthode des k-plus-proches-voisins associe la donnée $\boldsymbol{x}$ à la classe ω_i majoritairement représentée parmi ses k voisins



Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

Exemple de classification à 3 classes (rouge, vert, bleu), où chaque cluster de 20 points suit une loi normale de covariance identité.

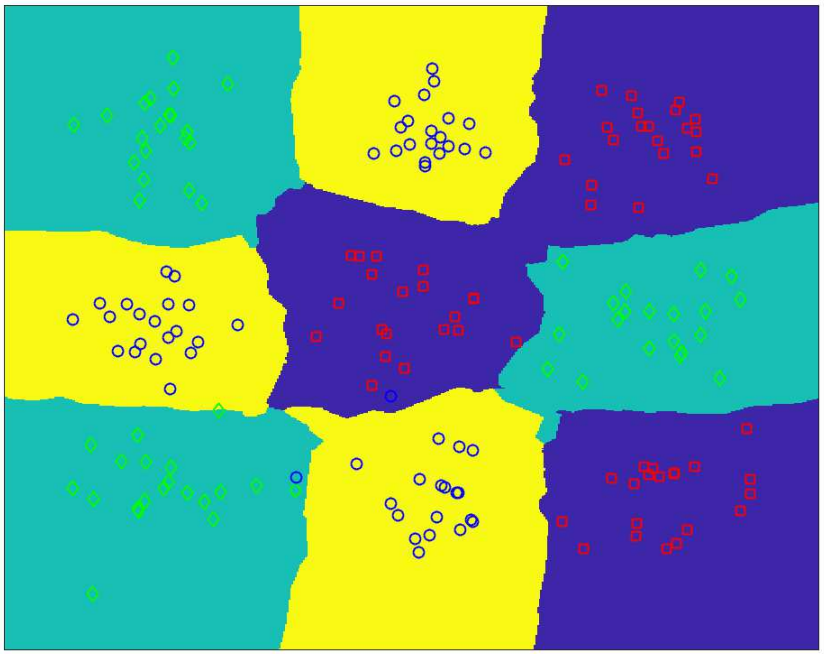


Estimation non-paramétrique de densité

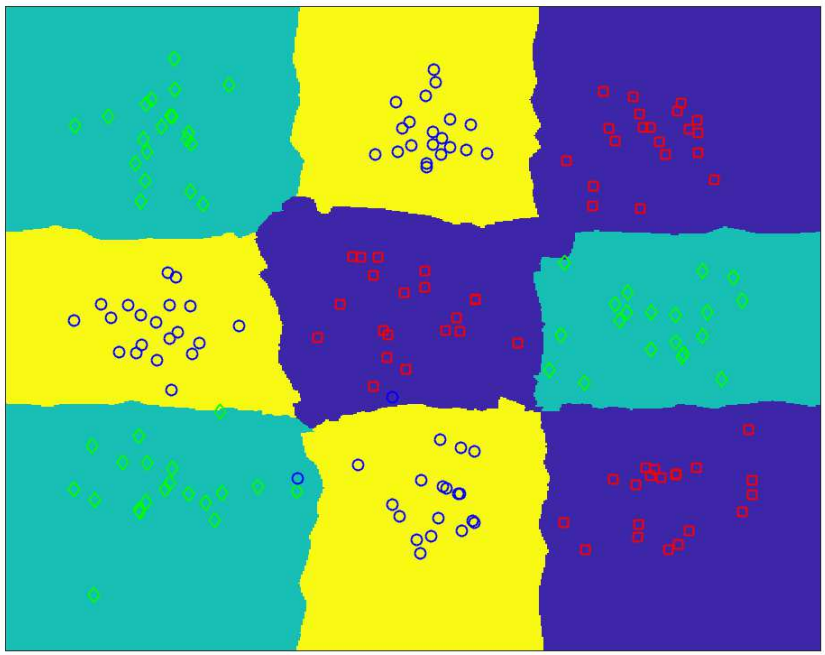
k-plus-proches-voisins : application à la classification

Exemple de classification à 3 classes (rouge, vert, bleu), où chaque cluster de 20 points suit une loi normale de covariance identité.

$k = 1$



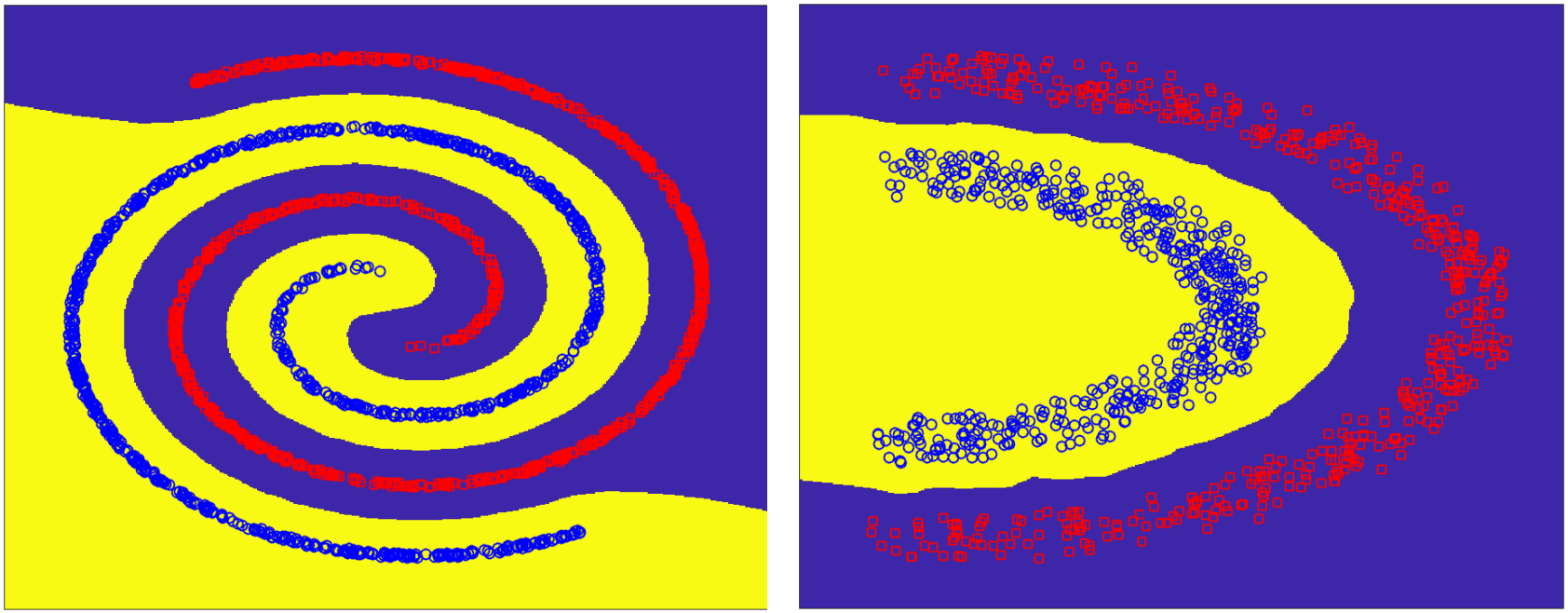
$k = 10$



Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

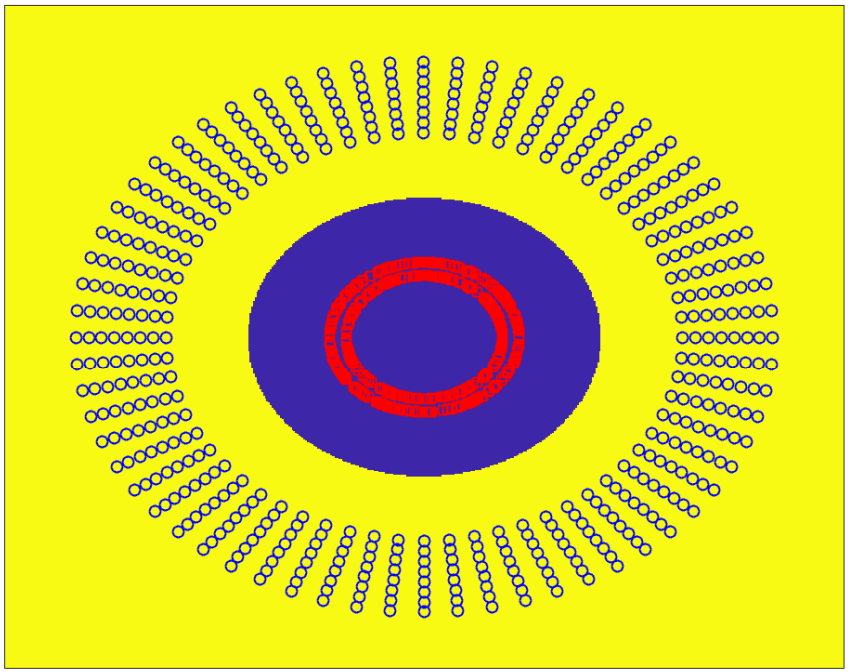
Résultats de la classification ($k = 1$)

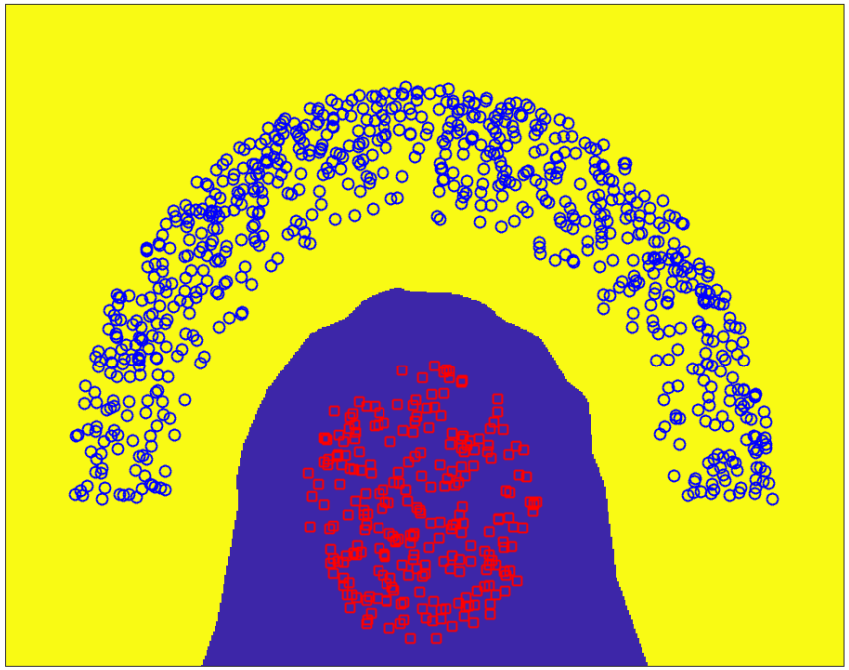


Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

Résultats de la classification ($k = 1$)

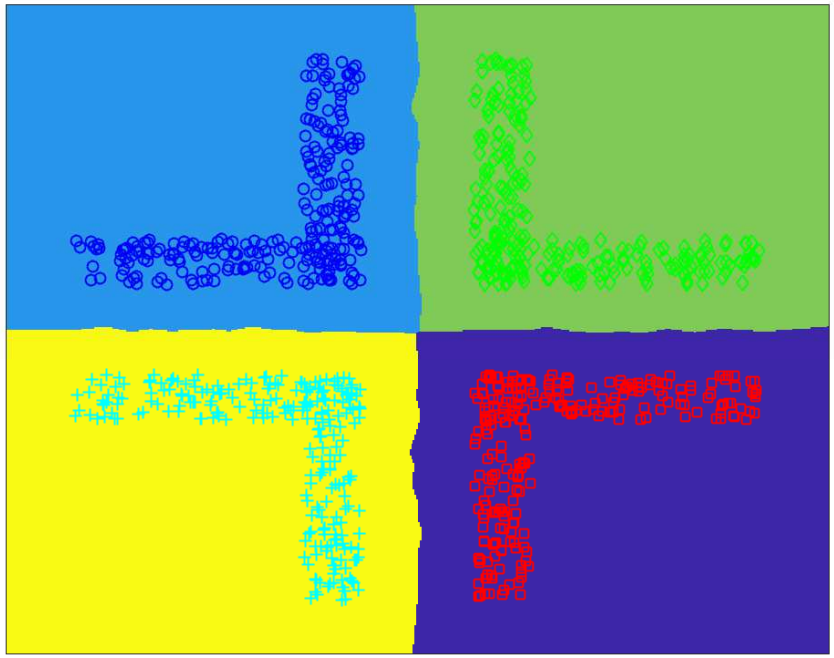


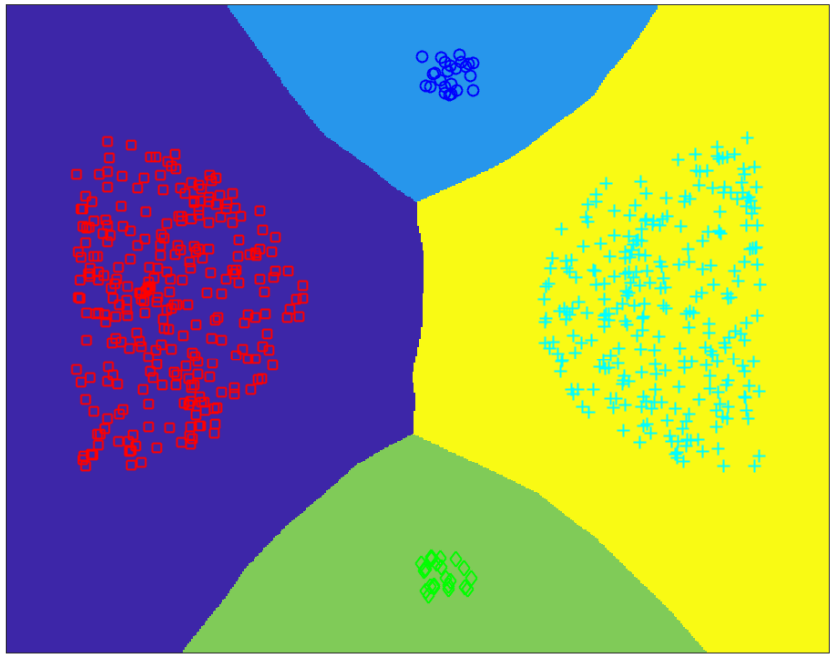


Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

Résultats de la classification ($k = 1$)





Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

L'algorithme des k-plus-proches-voisins n'extrait pas d'information des données afin d'estimer les paramètres d'un modèle.
En conséquence, la phase d'apprentissage demande pas/peu de ressources contrairement à la mise en œuvre.

Avantages :

- ▷ mise en œuvre simple, et implantation parallèle possible
- ▷ performant puisque, pour $N \to +\infty$, on a :

$$P_{Bayes}(\text{erreur}) < P_{1PPV}(\text{erreur}) < 2P_{Bayes}(\text{erreur})$$

- ▷ capacité d'adaptation car utilise l'information locale

Inconvénients :

- ▷ nécessite une grande capacité de stockage
- ▷ classification lourde d'un point de vue calculatoire
- ▷ sensible à la malédiction de la dimensionnalité

Estimation non-paramétrique de densité

k-plus-proches-voisins : application à la classification

De nombreuses stratégies d'améliorations existent pour

- ▷ réduire la taille de la base d'apprentissage en exploitant sa redondance
  - — en supprimant les données mal-classées, donc ambigües
  - — en supprimant les données bien-classées, donc loin de la frontière
- ▷ réduire le temps de recherche des voisins
  - — en partitionnant l'espace en cellules (diagramme de Voronoï)
  - — en hiérarchisant les données sous forme d'arbre
- ▷ choisir la métrique utilisée pour le calcul de distance